Tilburg University

# Asymptotic confidence intervals for the length of the shortt under random censoring

Einmahl, J.H.J.; Beirlant, J.

*Published in:*
Statistica Neerlandica

*Publication date:*
1995



*Citation for published version (APA):*
Einmahl, J. H. J., & Beirlant, J. (1995). Asymptotic confidence intervals for the length of the shortt under random censoring. *Statistica Neerlandica*, *49*(1), 1-8.

# statistica neerlandica


## Asymptotic confidence intervals for the length of the short*t* under random censoring


J. Beirlant

J. H. J. Einmahl



Journal of the Netherlands Society for Statistics and Operations Research

Volume 49 nr. 1 - March 1995

ISSN 0039-0402

*Published by Blackwell Publishers*

BLACKWELL Publishers

Statistica Neerlandica (1995) Vol. 49, nr. 1, pp. 1–8

# Asymptotic confidence intervals for the length of the short$t$ under random censoring

J. Beirlant[1]

*Dept. of Mathematics, Catholic University Leuven, Celestijnenlaan 200B, 3001 Heverlee, Belgium*

J. H. J. Einmahl

*Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

Dedicated to the memory of Nico Willems.

A short$t$ of a one dimensional probability distribution is defined to be an interval which has at least probability $t$ and minimal length. The length of a short$t$ and its obvious estimator are significant measures of scale of a distribution and the corresponding random sample, respectively. In this note a non-parametric asymptotic confidence interval for the length of the (uniqueness is assumed) short$t$ is established in the random censorship from the right model. The estimator of the length of the short$t$ is based on the product-limit (PL) estimator of the unknown distribution function. The proof of the result mainly follows from an appropriate combination of the Glivenko-Cantelli theorem and the functional central limit theorem for the PL estimator.

*Key Words & Phrases:* confidence interval, length of short$t$, random censorship.

## 1 Introduction and main result

Let $X_1, X_2, \ldots, X_n$ be a random sample from a univariate distribution function (df) $F$. An outlier resistant scale estimator based on such a sample is defined as the length of a shortest closed interval containing at least fraction $t$ (short$t$) of the data. This estimator possesses many desirable robustness properties; in particular for the case $t = 1/2$, the asymptotic break-down point is 50%; see ROUSSEEUW and LEROY (1988) for more details. A functional (in $t$) central limit theorem for this estimator is established in GRÜBEL (1988), see also EINMAHL and MASON (1992). It turns out that the length of the short$t$ has the "good" $n^{-1/2}$ rate of convergence, whereas the most prominent *location* estimators based on the short$t$ have only a rate of $n^{-1/3}$; cf. ANDREWS et al. (1972) and KIM and POLLARD (1990).

[1]Research performed while the author was research fellow at the Eindhoven University of Technology.

Let $F$ be continuous and write

$$U(t) = \inf\{b - a : F(b) - F(a) \geq t\}, \quad 0 < t < 1, \tag{1}$$

for the theoretical counterpart of the length of the short$t$, i.e. for our parameter of interest. It is the purpose of this note to derive a simple asymptotic confidence interval for $U(t)$, in the more general case that the $X_i$, $1 \leq i \leq n$, are randomly censored from the right.

In order to be more explicit, let us introduce some notation. Let $X_1, \ldots, X_n$ be as above and let $Y_1, \ldots, Y_n$ be an independent random sample from a df $G$, which we also assume to be continuous. In the random censorship from the right model we observe the independent pairs $(Z_i, \delta_i)$, $1 \leq i \leq n$, where $Z_i = X_i \wedge Y_i$ and $\delta_i = 1_{\{X_i \leq Y_i\}}$. The df of the $Z_i$ is denoted with $H$ and is easily seen to be equal to $1 - (1 - F)(1 - G)$. The well-studied product-limit estimator $F_n$ of $F$ (see e.g. GILL, 1980, SHORACK and WELLNER, 1986, Chapter 7) is given by

$$F_n(x) = 1 - \prod_{Z_{i:n} \leq x} \left(1 - \frac{\delta_{i:n}}{n - i + 1}\right), \quad x \in \mathbb{R},$$

where $Z_{1:n} \leq \cdots \leq Z_{n:n}$ are the order statistics of the $Z_i$ and $\delta_{i:n}$ are the corresponding $\delta$'s. Observe that trivially $F_n(x) = F_n(Z_{n:n})$ for $x > Z_{n:n}$. For $0 < t < 1$, let $U_n(t)$ be the empirical analogue of $U(t)$ based on $F_n$, i.e.

$$U_n(t) = \inf\{b - a : F_n(b) - F_n(a-) \geq t\}. \tag{2}$$

Write $[l_n, r_n]$ for the almost surely unique random interval corresponding to $U_n(t)$.

We also introduce the following empirical (sub-) df's:

$$H_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(Z_i), \quad x \in \mathbb{R},$$

$$H_n^1(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_i 1_{(-\infty, x]}(Z_i), \quad x \in \mathbb{R},$$

and write

$$D_n(x) = \int_{-\infty}^{x} \frac{1}{(1 - H_n(u-))^2} \, dH_n^1(u), \quad x < Z_{n:n}.$$

Furthermore, set

$$\hat{\sigma} = \{(1 - F_n(r_n))^2 (D_n(r_n) - D_n(l_n)) + t^2 D_n(l_n)\}^{1/2}$$

and let $c := c(\alpha)$ denote the $(1 - \alpha/2)$-th quantile of the standard normal df. In order to establish our result we need the following unimodality condition on $F$:

$F$ has a density $f$ which is positive and continuous on its support $(\beta, \gamma)$, $-\infty \leq \beta < \gamma \leq \infty$, strictly increasing on $(\beta, \eta]$ and strictly decreasing on $[\eta, \gamma)$ for some $\eta \in [\beta, \gamma)$. (3)

Let $[l, r]$ be the now uniquely defined interval corresponding to $U(t)$.

THEOREM. *Let* $0 < t < 1$ *be fixed, assume that* (3) *holds and that* $H(r) < 1$. *Then for any* $0 < \alpha < 1$

$$\lim_{n \to \infty} P\left(U_n\left(t - \frac{c\hat{\sigma}}{n^{1/2}}\right) < U(t) < U_n\left(t + \frac{c\hat{\sigma}}{n^{1/2}}\right)\right) = 1 - \alpha.$$

## 2 Proof of the result

For the proof we need the Glivenko-Cantelli theorem and the functional central limit theorem for $F_n$.

FACT 1. (*See e.g.* WANG, 1987 *or* STUTE *and* WANG, 1993. *As* $n \to \infty$

$$\sup_{x \leq Z_{n:n}} |F_n(x) - F(x)| \to_P 0.$$

A number of consequences of Fact 1 are stated in the next corollary. In the remainder of this proof $I$ denotes a closed interval $[a, b]$. For a function $g$ with left-hand limits, write

$$g(I) = g(b) - g(a-).$$

Denoting with $|I|$ the length of $I$ we define

$$\tilde{F}_n(y) = \sup_{\substack{|I| \leq y \\ I \subset (-\infty, Z_{n:n}]}} F_n(I),$$

$$\tilde{F}_{(n)}(y) = \sup_{\substack{|I| \leq y \\ I \subset (-\infty, Z_{n:n}]}} F(I).$$

COROLLARY 1. *Under the conditions of the Theorem we have for small enough* $\varepsilon > 0$, *as* $n \to \infty$,

$$\sup_{I \subset (-\infty, Z_{n:n}]} |F_n(I) - F(I)| \to_P 0, \tag{4}$$

$$\sup_{|y - U(t)| \leq \varepsilon} |\tilde{F}_n(y) - \tilde{F}_{(n)}(y)| \to_P 0, \tag{5}$$

$$U_n(t) \to_P U(t), \tag{6}$$

$$\sup_{x \leq Z_{n:n} - U_n(t)} |F_n([x, x + U_n(t)]) - F([x, x + U(t)])| \to_P 0, \tag{7}$$

$$l_n \to_P l \quad \text{and} \quad r_n \to_P r. \tag{8}$$

PROOF: The statement in (4) trivially follows from Fact 1 and the assertion in (5) follows immediately from (4). Write

$$\tilde{F}(y) = \sup_{|I| \leq y} F(I)$$

and note that $U$ is continuous and strictly increasing on (0, 1) (because of (3)) and that $\tilde{F}$ is its inverse. Now (5) implies that for small $\delta > 0$

$$P(\tilde{F}_n(U(t-\delta)) < t \leq \tilde{F}_n(U(t+\delta))) \to 1 \ (n \to \infty). \tag{9}$$

Observe that

$$U_n(t) = \inf\{y : \tilde{F}_n(y) \geq t\}, \quad 0 < t < 1,$$

and hence from (9) we have

$$P(U(t-\delta) < U_n(t) \leq U(t+\delta)) \to 1 \ (n \to \infty),$$

which yields (6).

The proof of (7) follows from the fact that its left-hand side is less than or equal to

$$\sup_{x \leq Z_{n:n} - U_n(t)} |F_n([x, x + U_n(t)]) - F([x, x + U_n(t)])| \\ + \sup_{x \leq Z_{n:n} - U_n(t)} |F([x, x + U_n(t)]) - F([x, x + U(t)])|,$$

in combination with (4), (6) and the uniform continuity of $F$. Finally, assertion (8) is a consequence of (7) and (6); see e.g. KIM and POLLARD, 1990, p. 208). □

To present the functional central limit theorem for $F_n$, which we state in an almost sure construction setting, write

$$H^1(x) = P(Z_i \leq x, \delta_i = 1) = \int_{-\infty}^{x} (1 - G(u))\, dF(u), \quad x \in \mathbb{R},$$

$$D(x) = \int_{-\infty}^{x} \frac{1}{(1 - H(u))^2}\, dH^1(u), \quad x < \sup\{y : H(y) < 1\},$$

and

$$\alpha_n(x) = n^{1/2}(F_n(x) - F(x)), \quad x \in \mathbb{R}.$$

FACT 2. (*See e.g. SHORACK and WELLNER,* 1986, *p.* 308.) *Let* $R \in \mathbb{R}$ *with* $H(R) < 1$. *Under the conditions of the Theorem exists a sequence of processes* $\{\tilde{\alpha}_n\}_{n=1}^{\infty}$, *with* $\tilde{\alpha}_n \stackrel{d}{=} \alpha_n$, *and a standard Wiener process* $W$ *such that as* $n \to \infty$

$$\sup_{x \leq R} |\tilde{\alpha}_n(x) - (1 - F(x))W(D(x))| \to 0 \text{ a.s.} \tag{10}$$

REMARK 1. Throughout we will choose $R > r$.

REMARK 2. To prove our result we will proceed on the probability space on which (10) holds. Without confusion, we shall henceforth drop the symbol $^{-}$ from the notation.

Write $V(x) = (1 - F(x))W(D(x))$.

COROLLARY 2. *As* $n \to \infty$

$$\sup_{I \subset (-\infty, R]} |\alpha_n(I) - V(I)| \to 0 \text{ a.s.}$$

Define $\tilde{\alpha}_n(y) = n^{1/2}(\sup_{|I| \le y} F_n(I) - \tilde{F}(y))$; hence $\tilde{\alpha}_n(U(t)) = n^{1/2}(\sup_{|I| \le U(t)} F_n(I) - t)$.

PROPOSITION. *Under the conditions of the Theorem we have as* $n \to \infty$

$$\tilde{\alpha}_n(U(t)) \to V([l, r]) \text{ a.s.} \tag{11}$$

REMARK 3. It is readily checked that $V([l, r])$ is a centered normal random variable with variance $(1 - F(r))^2(D(r) - D(l)) + t^2 D(l) =: \sigma^2$.

PROOF: Define

$$\tilde{F}_{n,R}(y) = \sup_{\substack{|I| \le y \\ I \subset (-\infty, R]}} F_n(I),$$

$$\tilde{F}_R(y) = \sup_{\substack{|I| \le y \\ I \subset (-\infty, R]}} F(I),$$

and

$$\tilde{\alpha}_{n,R}(y) = n^{1/2}(\tilde{F}_{n,R}(y) - \tilde{F}_R(y)).$$

Observe that because of Remark 1 we have $\tilde{F}_R(U(t)) = \tilde{F}(U(t)) = t$. Also

$$\lim_{n \to \infty} P(\tilde{F}_{(n)}(U(t)) = \tilde{F}(U(t))) = 1. \tag{12}$$

Furthermore, we have for the empirical counterparts of these quantities that $\tilde{F}_n(U(t)) = \sup_{|I| \le U(t)} F_n(I)$, and because of (4), (5) and (12)

$$\lim_{n \to \infty} P(\tilde{F}_n(U(t)) = \tilde{F}_{n,R}(U(t))) = 1.$$

Therefore it suffices to prove (11) with $\tilde{\alpha}_n(U(t))$ replaced by $\tilde{\alpha}_{n,R}(U(t))$. The proof of this will be given along the lines of the proof of Proposition 3.1 in EINMAHL and MASON (1992).

First observe that

$$V([l,r]) - \tilde{\alpha}_{n,R}(U(t)) \leqslant V([l,r]) - n^{1/2}(F_n([l,r]) - F([l,r])).$$

Hence by Corollary 2

$$\limsup_{n\to\infty} V([l,]) \dot{-}_{n,R}(U(t)) \leqslant 0 \text{ a.s.}$$

We also have

$$\tilde{\alpha}_{n,R}(U(t)) - V([l,r]) \leq \left\{ n^{1/2}\left( \sup_{\substack{|I|\leq U(t)\\ I\subset(-\infty,R]\\ t-n^{-1/4}<F(I)\leq t}} F_n(I) - t \right) - V([l,r]) \right\}$$
$$\vee \left\{ n^{1/2}\left( \sup_{\substack{I\subset(-\infty,R]\\ F(I)\leq t-n^{-1/4}}} F_n(I) - t \right) - V([l,r]) \right\}. \tag{13}$$

The second term on the right of (13) is less than or equal to

$$\sup_{\substack{I\subset(-\infty,R]\\ F(I)\leq t}} n^{1/2}(F_n(I) - F(I)) + |V([l,r])| - n^{1/4}$$
$$\leq \sup_{I\subset(-\infty,R]} |\alpha_n(I) - V(I)| + 2 \sup_{I\subset(-\infty,R]} |V(I)| - n^{1/4},$$

which by Corollary 2 and the fact that $H(R) < 1$, converges to $-\infty$ almost surely. The first term on the right of (13) is less than or equal to

$$n^{1/2} \sup_{\substack{|I|\leq U(t)\\ I\subset(-\infty,R]\\ t-n^{-1/4}<F(I)\leq t}} (F_n(I) - F(I)) - V([l,r])$$
$$\leqslant \sup_{I\subset(-\infty,R]} |\alpha_n(I) - V(I)| + \left( \sup_{\substack{|I|\leq U(t)\\ I\subset(-\infty,R]\\ t-n^{-1/4}<F(I)\leq t}} V(I) \right) - V([l,r]). \tag{14}$$

From again Corollary 2 and condition (3) in conjunction with the continuity of $V$ we see that the right-hand side of (14) converges to 0 almost surely. □

REMARK 4. In principle

$$\tilde{\alpha}_{n,R}(U(t)) \to V([l,r]) \text{ a.s. } (n\to\infty),$$

follows readily from Proposition 8 in GRÜBEL (1988) and Fact 2, in conjunction with the functional delta method (see e.g. GILL, 1989). We have chosen for the present technique, however, because then the conditions on the density $f$ occur to be milder, especially the differentiability of the density is not required.

We finally need the following

LEMMA. *Under the conditions of the Theorem we have as $n \to \infty$*

$$\hat{\sigma} \longrightarrow_p \sigma.$$

PROOF: Easy; based on Fact 1, (8) and the well known fact that

$$\sup_{x \leq R} |D_n(x) - D(x)| \rightarrow_p 0 \ (n \to \infty). \qquad \square$$

PROOF OF THE THEOREM: From the Proposition and the Lemma it is immediate that, as $n \to \infty$, $\tilde{\alpha}_n(U(t))/\hat{\sigma}$ converges weakly to a standard normal random variable, implying that

$$P\left(U_n\left(t - \frac{c\hat{\sigma}}{n^{1/2}}\right) \leq U(t) < U_n\left(t + \frac{c\hat{\sigma}}{n^{1/2}}\right)\right)$$

$$= P\left(t - \frac{c\hat{\sigma}}{n^{1/2}} \leq \sup_{|I| \leq U(t)} F_n(I) < t + \frac{c\hat{\sigma}}{n^{1/2}}\right)$$

$$= P(-c \leq \tilde{\alpha}_n(U(t))/\hat{\sigma} < c) \to 1 - \alpha.$$

A little reflection shows that

$$P\left(U_n\left(t - \frac{c\hat{\sigma}}{n^{1/2}}\right) = U(t)\right) = 0.$$

This completes the proof. $\square$

REMARK 5. Apart from technical details, associated with the random censorship from the right model, it is clear that in major parts of the proof the fact that $F$ is estimated by the product-limit estimator $F_n$ plays no essential role. Hence similar results to the one in this paper can be derived along the same lines for other weakly convergent estimators of a distribution function $F$.

## References

ANDREWS, D. F., P. J. BICKEL, F. R. HAMPEL, P. J. HUBER, W. H. ROGERS and J. W. TUKEY (1972), *Robust estimates of location*, Princeton University Press, Princeton, N.J.

EINMAHL, J. H. J. and D. M. MASON (1992), Generalized quantile processes, *The Annals of Statistics* 20, 1062–1078.

GILL, R. D. (1980), *Censoring and stochastic integrals*, Mathematical Centre Tract 124, CWI, Amsterdam.

GILL, R. D. (1989), Non- and semiparametric maximum likelihood estimators and the von Mises method (Part I), *Scandinavian Journal of Statistics* 16, 97–128.

GRÜBEL, R. (1988), The length of the shorth, *The Annals of Statistics* 16, 619–628.

KIM, J. and D. POLLARD (1990), Cube root asymptotics, *The Annals of Statistics* 18, 191–219.

ROUSSEEUW, P. and A. LEROY (1988), A robust scale estimator based on the shortest half, *Statistica Neerlandica* 42, 103–116.

SHORACK, G. R. and J. A. WELLNER (1986), *Empirical processes with applications to statistics*, Wiley, New York.

STUTE, W. and J.-L. WANG (1993), The strong law under random censorship, *The Annals of Statistics* 21, 1591–1607.

WANG, J.-G. (1987), A note on the uniform consistency of the Kaplan-Meier estimator, *The Annals of Statistics* 15, 1313–1316.

Received: February 1993. Revised: February 1994.